STANDARDIZED POPULATION COMPARISONS FOR A SAMPLE SURVEY

Rick L. Williams and Lisa Morrissey LaVange
Research Triangle Institute

1. Introduction

An integral part of many analyses is the assessment of differences between subgroups with respect to a population characteristics. When estimating such differences it is often necessary to take into account other factors that may not be of interest themseleves but that could cloud the effect being studied. An extraneous variable that is associated with both the population characteristics being measured and the study factor (subgroups being compared) is a confounding variable. If its distribution varies across levels of the study factor then failure to control for the variable will confound the estimate of the study factor effect.

The most straightforward method of controlling for a confounding variable is to divide the population into separate control groups according to values of the confounding variable (or variables). The effect of the study factor is then determined within each control group. Analyses of this type are typically referred to as stratified analyses.

Stratified analyses require separate estimation of the effects within each control group. This may prove impractical if the number of control groups becomes large. The sample size within individual control groups may become too small to be of value. Also, the researcher must form his/her conclusions based on a multitude of estimates. For these reasons it is often desirable to summarize the relationship between the study factor and the outcome. Standardization is a common method for summarizing the results of a stratified analysis. A standardized estimate is a weighted average of the control group specific estimates.

The decision to standardize or not depends on the question posed. If it is desired to determine the actual difference between two study group populations as they currently exist, the non-standardized difference between the overall study group means is appropriate. On the other hand, the explanation of the underlying determinants of a study factor effect may require standardization.

The remainder of this paper will focus on estimating standardized or adjusted means. The simple arithmetic difference between two study group adjusted means will be the effect measure of central discussion. Both direct standardization and regression standardization will be discussed.

Standardization is not a new topic. This paper elucidates some aspects of standardization when analyzing sample survey data. The population value of interest is first defined; then a method of estimating the population value with sample survey data is presented. Inferential methods concerning the population values are discussed. A general unequally weighted sample is assumed.

2. Methods of Standardization

2.1 Direct Standardization

2.1.1 General

As a simple example consider the data for the population given in Table 1. In this example, the confounding variable has been grouped into three control groups, and the study factor takes on two levels. Here $\bar{Y}_{ij}$ is the population mean for the N_{ij} individuals in the ith study group and the jth control group. The overall differ-ence between the two study groups is 5.18 ($= \bar{Y}_2 - \bar{Y}_1$), the differences within control group are all 3.00 ($= \bar{Y}_{1j} - \bar{Y}_{2j}$ for all j). This indicates that the relationship between the study factor and the outcome is confounded with the control variable. Direct standardization would control for the type of confounding in the above example by forming, for each study group, the directly standardized population means

$$\bar{Y}_i^* = \sum_j p_j \bar{Y}_{ij} \qquad (2.1)$$

where $p_j > 0$ and $\sum_j p_j = 1$.

The distribution $\{p_j\}$ is the standardizing distribution. The choice of a standardizing distribution will be discussed later. The directly standardized means are the population values of interest.

The difference between the two standardized means is

$$\bar{Y}_2^* - \bar{Y}_1^* = \sum_j p_j (\bar{Y}_{2j} - \bar{Y}_{1j}). \qquad (2.2)$$

Two cases arise which require separate discussion. The first is when the relationship between the study factor and the confounding variable is additive (i.e. no interaction is present) in the population. The second is when the relationship is nonadditive (i.e. interaction is present).

Additive Case

The additive case implies that a common difference d ($= \bar{Y}_{2j} - \bar{Y}_{1j}$ for all j) is present in in the population for each control group. Thus, in the population,

$$\begin{aligned} \bar{Y}_2^* - \bar{Y}_1^* &= \sum_j p_j (\bar{Y}_{2j} - \bar{Y}_{1j}) \\ &= \sum_j p_j \, d \\ &= d \quad . \end{aligned} \qquad (2.3)$$

For the data in Table 1, d = 3. Notice that (2.3) shows that in the additive case the contrast between the two adjusted means is independent of the standardizing distribution $\{p_j\}$.

Thus, the effect of the study factor, as measured by the difference between the two adjusted means, is free of the effect of the confounding variable. This situation was discussed in detail by Kalton (1968).

Non-Additive Case

On the other hand, in the non-additive case, the study effect or difference varies across the control groups. In this case,

$$\bar{Y}_2^* - \bar{Y}_1^* = \Sigma\, p_j\, (\bar{Y}_{2j} - \bar{Y}_{1j})$$
$$= \Sigma_j\, p_j\, d_j \qquad (2.4)$$

where d_j is the control group specific difference between the two study groups. The difference between the two standardized means is equivalent to a weighted average of the control group specific effects. This is the case for the population shown in Table 2. The difference between the two study group means increases across the control groups. (1.0 vs. 2.0 vs. 6.0). The non-additivity between the study factors and the confounding variable prevents the estimation of a study factor effect free of the influence of the confounder.

As noted by Kitagawa (1964) and Little (1982) it may still be desirable to summarize the control group specific study factor effects through standardization. When the interactions are not severe, a standardized measure of effect may provide a meaningful summary which is more readily interpretable than the results of a stratified analysis. Some information is sacrificed in such a summary. When summarizing over nuisance variables, the loss in information may be unimportant. However, if the interactions are severe, the loss of information through standardization may be unacceptable. In fact, contradictory conclusions may be obtained by choosing two different standardizing populations.

For the second example, the unadjusted difference between the two study groups is 6.18 ($= \bar{Y}_2 - \bar{Y}_1$). The difference adjusting to the marginal distribution of the confounding variable for the combined study groups is 3.09 ($= \bar{Y}_2^* - \bar{Y}_1^*$). The standardized effect mitigates the influence of the differential population distributions.

2.1.2 Standardizing Distribution

As shown in equation 2.1., the directly standardized mean for a particular study group is a weighted average of the control group specific means. The weights, $\{p_j\}$, represent the proportion of the hypothetical population in each control group.

It was also shown (see equation 2.3) that in the case of additive control group and study factor effects, the standardized difference is independent of the standardizing distribution. In this case, the standardizing distribution should generally be taken to maximize the precision of the study factor effect. Optimal selections for the standardizing distribution are presented by Kalton (1968) for the additive case.

When the control group and study factor effects are not additive, the choice of a standardizing distribution becomes more critical since the exact conclusions drawn depend upon this distribution. In this situation, we have found that the marginal distribution obtained by combining the study groups usually performs well. Using the marginal distribution insures that the standardized difference is interpolated at the "center" of the data rather than extrapolated from an extreme point.

2.2 Regression Standardization

2.2.1 General

Regression analysis is another method that can be used to estimate the effect of the study factor while controlling for extraneous variables. The direct standardization approach is a special case of regression standardization. The regression approach is analogous to the analysis of covariance (e.g. Snedecor and Cochran, 1967).

Regression standardization assumes that the response can be predicted by a linear model involving the study factor and the confounding variables. The model is used to predict adjusted or standardized means for the study groups by assuming each group has the same values of the confounding variables. Comparison of the adjusted means partially removes the effect of the confounding variables.

In order to formulate the regression model for the population, the following definitions are made:

$\underset{\sim}{Y}$ = population vector of the outcome measure,

$\underset{\sim}{\gamma}$ = vector of model parameters corresponding to the study factor,

$\underset{\sim}{Z}$ = population design matrix for the study factor,

$\underset{\sim}{\beta}$ = vector of model parameters corresponding to the confounding variables and their interactions with the study factor,

$\underset{\sim}{X}$ = population design matrix corresponding to $\underset{\sim}{\beta}$.

The vectors $\underset{\sim}{\gamma}$ and $\underset{\sim}{\beta}$ are futher defined by the relationship

$$\begin{vmatrix} \underset{\sim}{\gamma} \\ \underset{\sim}{\beta} \end{vmatrix} = \{[\underset{\sim}{Z},\underset{\sim}{X}]'\,[\underset{\sim}{Z},\underset{\sim}{X}]\}^{-1}\,[\underset{\sim}{Z},\underset{\sim}{X}]'\,\underset{\sim}{Y} \quad .$$

The regression model is then given by

$$E(\underset{\sim}{Y}) = \underset{\sim}{Z}\,\underset{\sim}{\gamma} + \underset{\sim}{X}\,\underset{\sim}{\beta}.$$

The adjusted population mean for study group-i is

$$\bar{Y}_i^* = \underset{\sim}{Z}_i\,\underset{\sim}{\gamma} + \underset{\sim}{X}_i^*\,\underset{\sim}{\beta}.$$

The vector $\underset{\sim}{Z}_i$ contains the linear transformation of $\underset{\sim}{\gamma}$ that yields the intercept for study

group-i. Usually, $\underset{\sim}{Z}_i$ will have a one in the ith position and zeros elsewhere. The vector $\underset{\sim}{X}^*_i$ contains the standardizing values of the confounding variables used in common across the study groups. The structure of $\underset{\sim}{X}^*_i$, but not the standardizing values, will vary from subgroup to subgroup if the model contains interactions between the study factor and any of the confounding variables. Regression standardization is equivalent to direct standardization when a fully interactive regression model is assumed and all the variables are categorical.

With the interactions between the study factor and the confounding variables in the model, care must be taken in interpreting the difference between two adjusted means. The adjusted difference is a function of the standardizing population vector ($\underset{\sim}{X}^*_i$). This may produce a useful condensation of the data if the interactions are not too severe and the standardizing distribution is chosen appropriately. As was the case for direct standardization, it is our experience that standardizing to the mean of the confounding variables over all of the study groups is usually suitable. This insures that the standardized difference is taken at the center of the data rather than at some extreme point.

When the interactions between the study factor and the confounding variables are not present, the selection of the standardizing distribution is less crucial since the standardized difference is independent of the distribution. However, it is useful to make sure that the adjusted means are interpretable and are reasonable values. Using the marginal mean of the confounding variables over all the study groups as the standardizing distribution, X^*_i, will usually satisfy this requirement.

2.2.2 Estimation

The estimation of regression standardized means from sample survey data is presented in this section. The following definitions are needed:

$\underset{\sim}{y}$ = outcome measurement vector for the sample subjects,

$\underset{\sim}{z}$ = study factor design matrix for the sample,

$\underset{\sim}{x}$ = design matrix of the confounding variables and their interactions with the study factor for the sample subjects,

$\underset{\sim}{w}$ = diagonal matrix of the sampling weights.

With these definitions, the population parameter vectors can be estimated with

$$\begin{vmatrix} \hat{\underset{\sim}{y}} \\ \hat{\underset{\sim}{\beta}} \end{vmatrix} = \{[\underset{\sim}{z},\underset{\sim}{x}]'\underset{\sim}{w}[\underset{\sim}{z},\underset{\sim}{w}]\}^{-1}\ [\underset{\sim}{z},\underset{\sim}{x}]'\underset{\sim}{w}\underset{\sim}{y}$$

For most sample designs the variance covariance matrix of $[\hat{\underset{\sim}{y}}, \hat{\underset{\sim}{\beta}}]'$ can be estimated using the Taylor series linearization, balance repeated replication or the jackknife method. Let $\hat{\underset{\sim}{V}}$ be the estimated covariance matrix of $[\hat{\underset{\sim}{y}}, \hat{\underset{\sim}{\beta}}]'$. At this point hypothesis concerning the model parameters can be tested via large sample Wald statistics. The model may be reduced to remove any non-significant terms.

The estimated standardized mean for study group-i is

$$\bar{y}^*_i = \underset{\sim}{Z}_i\ \hat{\underset{\sim}{y}} + \underset{\sim}{X}^*_i\ \hat{\underset{\sim}{\beta}}$$

$$= g_i \begin{vmatrix} \hat{\underset{\sim}{y}} \\ \hat{\underset{\sim}{\beta}} \end{vmatrix}$$

where g_i is the linear transformation of $[\hat{\underset{\sim}{y}}, \hat{\underset{\sim}{\beta}}]'$ that yields $\bar{y}^*_i$. Further letting $\bar{\underset{\sim}{y}}^*$ be the column vector of the estimated regression adjusted means and assuming that there are s study groups yields

$$\bar{\underset{\sim}{y}}^* = \begin{vmatrix} \bar{y}^*_1 \\ \cdot \\ \cdot \\ \cdot \\ \bar{y}^*_s \end{vmatrix}$$

$$= \underset{\sim}{G} \begin{vmatrix} \hat{\underset{\sim}{y}} \\ \hat{\underset{\sim}{\beta}} \end{vmatrix}$$

where $\underset{\sim}{G} = [g_1', \ldots, g_s']'$. Hence, the estimated covariance matrix of $\bar{\underset{\sim}{y}}^*$ is

$$\hat{V}(\bar{\underset{\sim}{y}}^*) = \hat{\underset{\sim}{V}}^*$$

$$= \underset{\sim}{G}\ \hat{\underset{\sim}{V}}\ \underset{\sim}{G}'.$$

As noted earlier, direct standardization is a special case of regression standardization. Hence, by estimating the appropriate fully interactive model directly standardized mean estimates can be produced.

Hypotheses concerning the previously defined population values can be tested using their sample design based estimates. For example, letting $\bar{\underset{\sim}{Y}}^* = [\bar{Y}^*_1, \ldots, \bar{Y}^*_s]'$ be the vector of population values for the adjusted means, a linear hypothesis concerning these means may be stated as

$$H_0: \ \underset{\sim}{C}\ \bar{\underset{\sim}{Y}}^* = \underset{\sim}{0}$$

vs.

$$H_A: \ \underset{\sim}{C}\ \bar{\underset{\sim}{Y}}^* \neq \underset{\sim}{0}.$$

where $\underset{\sim}{C}$ is a matrix of contrasts and $\underset{\sim}{0}$ is a null vector. A large sample Wald test of this hypothesis would use the quadratic form

$$Q = [\underset{\sim}{C}\ \bar{\underset{\sim}{y}}^*]'\ [\underset{\sim}{C}\ \hat{\underset{\sim}{V}}^*\ \underset{\sim}{C}']^{-1}\ [\underset{\sim}{C}\ \bar{\underset{\sim}{y}}^*].$$

Under the null hypothesis, Q is asymptotically distributed as a chi-square random variable with rank ($\underset{\sim}{C}$) degress of freedom.

2.3 Discussion

The regression standardization approach offers several advantages over the direct method. One of these is that the potential confounders can be tested for significant association with the outcome measure. Non-significant terms can be dropped from the model so that more precise estimates of the adjusted means can be obtained. In addition, regression standardization does not require continuous confounding variables to be categorized. They may be included in the model as polynomial effects.

Direct standardization requires separte cell estimates for the complete cross-classification of all the confounding and study variables. This limits the number of confounding variables that can be controlled due to small sample sizes in each cell of the cross-classification. With the regression approach, the complete interaction of all the confounding variables need not be included in the model.

3. Example

Data from the National Medical Care Utilization and Expenditure Survey (NMCUES) will be used to provide a brief example. The State Medicaid Household Survey (SMHS) component of NMCUES collected data from a sample of 1,000 Medicaid families from each of the States of California, Michigan, New York and Texas. A multi-stage cluster sample was used to select the Medicaid families in each State. Data were collected on all medical care events during 1980 from the survey subjects.

An important part of the SMHS data analysis is the comparison of health care utilization rates among the four states. Because the Medicaid enrollees in each state differ considerably with respect to various extraneous factors believed to affect health care utilization, it was necessary to control for possible confounding due to these factors before making these comparisons. The possible confounding variable included in the model were:

Age	Health Status
Race	SMSA vs Non-SMSA
Sex	Income
Hispanic Origin	Education.

Separate comparison of the state specific utilization rates were done for the four Medicaid aid categories:

SSI Blind or Disabled
SSI Aged
Aid to Families with Dependent Children (AFDC)
State Only.

Standardization models were fit to the data for each of the aid categories and for several types of utilization. The models initially contained the state study factor variable, the previously listed confounding variables and the interactions of state with the confounding variables. The models were reduced to remove effects that did not significantly affect the utilization measure under consideration.

Table 3 presents the estimated adjusted and unadjusted mean number of physician visits per person for 1980 by State and aid category. The estimated standard error of each value is also included. Notice that the adjusted and unadjusted means are quite similar. This indicates that the state comparisons are not confounded with the extraneous variables.

REFERENCES

Kalton, G. (1968), "Standardization: A Technique to Control for Extraneous Variables," Applied Statistics, Vol. 23.

Kitagawa, E. M. (1964), "Standardized Comparisons in Population Research," Demography, Vol. 1.

Little, R.J.A. (1982), "Direct Standardization: A Tool for Teaching Linear Models for Unbalanced Data," American Statistician, Vol. 36, No.1.

Snedecor, G. W., W. G. Cochran (1967), Statistical Methods, Sixth Edition. The Iowa State University Press, Ames, Iowa.

Table 1. Example of Confounding Without Interaction

Study Groups	Control Groups 1	2	3	
1	$\bar{Y}_{11}$=2.00	$\bar{Y}_{12}$=4.00	$\bar{Y}_{13}$=8.00	$\bar{Y}_{1}$=3.64
	N_{11}=150	N_{12}=75	N_{13}=50	N_{1}=275
2	$\bar{Y}_{21}$=5.00	$\bar{Y}_{22}$=7.00	$\bar{Y}_{23}$=11.00	$\bar{Y}_{2}$=8.82
	N_{21}=50	N_{22}=75	N_{23}=150	N_{2}=275

Table 2. Example of Confounding with Interaction

Study Groups	Control Groups 1	2	3	
1	$\bar{Y}_{11}$=2.00	$\bar{Y}_{12}$=4.00	$\bar{Y}_{13}$=8.00	$\bar{Y}_{1}$=3.64
	N_{11}=150	N_{12}=75	N_{13}=50	N_{1}=275
2	$\bar{Y}_{21}$=3.00	$\bar{Y}_{22}$=6.00	$\bar{Y}_{23}$=14.00	$\bar{Y}_{2}$=9.82
	N_{21}=50	N_{22}=75	N_{23}=150	N_{2}=275

Table 3. Adjusted and Unadjusted Mean Number of Physician Visits Per Person by Aid Category and State for 1980

Aid Category/ State	Unadjusted Mean	Unadjusted Standard Error	Adjusted Mean	Adjusted Standard Error
SSI Blind or Disabled				
CA	13.92	1.19	13.93	1.10
MI	10.33	.65	10.10	.67
NY	15.97	1.46	16.47	1.35
TX	9.23	.74	8.52	.68
SSI Aged				
CA	11.40	.97	11.58	.93
MI	8.22	.47	8.55	.96
NY	10.21	.76	10.06	.71
TX	7.17	.47	6.78	.43
AFDC				
CA	4.20	.21	4.21	.21
MI	4.21	.23	4.35	.20
NY	5.01	.30	4.88	.27
TX	3.13	.18	3.44	.17
State Only*				
CA	6.46	.53	7.06	.56
MI	6.30	.84	8.01	1.29
NY	12.59	1.73	11.43	1.56

*Texas did not have a state only program.